Submitted, 2019 American Control Conference (ACC)
http://www.cds.caltech.edu/~murray/preprints/chmc19-acc_s.pdf

Robust Estimation Framework with Semantic Measurements

Karena X. Cai[1], Alexei Harvard, Richard M. Murray, Soon-Jo Chung

***Abstract*— Conventional simultaneous localization and mapping (SLAM) algorithms rely on geometric measurements and require loop-closure detections to correct for drift accumulated over a vehicle trajectory. Semantic measurements can add measurement redundancy and an alternative form of loop closure. We propose two different estimation algorithms that incorporate semantic measurements provided by vision-based object classifiers. An a priori map of regions where the objects can be detected is assumed. The first estimation framework is posed as a maximum-likelihood problem, where the likelihood function for semantic measurements is derived from the confusion matrices of the object classifiers. The second estimation framework is comprised of two parts: 1) a continuous-state estimation formulation that includes semantic measurements as a form of state constraints and 2) a discrete-state estimation formulation used to compute the certainty of object detection measurements using a Hidden Markov Model (HMM). The advantages of incorporating semantic measurements in these frameworks are demonstrated in numerical simulations. In particular, the proposed estimation algorithms improve upon the robustness and accuracy of conventional SLAM algorithms. Also, the certainty metric of object detection measurements derived from the HMM in our simulation are greater than the certainty levels provided by the confusion matrix in object classification algorithms.**

I. INTRODUCTION

The widespread availability of vision-based object classification tools has opened up the possibility of including semantic data into simultaneous localization and mapping (SLAM) algorithms. The data extracted from object-classifiers, which we refer to as semantic data, can be used to complement the more conventional nonlinear continuous measurements that are traditionally used in SLAM algorithms.

Semantic data can be modeled as binary measurements that have state-dependent probabilistic likelihood functions [1], [2], [3]. The probability of a positive detection measurement is modeled as an inverse-exponential function of the distance to the detected object in [1], meaning positive object detections occur with higher probability when the vehicle is close to the detected object. The authors in [4] compute the likelihood function of an object detection event as a function of the object classifier confusion matrix, and solve the coupled data association and estimation problem by iteratively solving an expectation-maximization problem. In these algorithms, however, the likelihood functions lack the ability to capture false positive or negative detections. A likelihood function that captures these types of errors is derived in [3], but requires additional assumptions on the probability of false positive detections generated by clutter.

Since object classifiers are prone to false positive and negative detections, an accurate estimation algorithm that integrates semantic data requires the rejection of false measurements. Loop closure events, which are the detection of returning to a previously visited location, are similar to object detection events, since they add a state constraint and the inclusion of false loop closure measurements can cause large errors in the state estimate [5].

Several algorithms have been developed to ensure the factor-graph formulation for SLAM problems are robust to loop-closure errors [5], [6], [7]. In particular, the covariance value associated with loop closure measurements, which are referred to as switchable constraint variables, are introduced into the optimization framework to improve robustness to false detections [7], [8]. In this paper, we extend such methods to robustly handle false object detection measurements as well. This requires deriving a method for computing the certainty of object detection measurements. The method we derive for computing this certainty metric leverages information from pose estimation.

Pose estimation has been shown to improve the accuracy of object classification algorithms [9], [10]. These methods take into account the motion profile taken when observing an object, but do not consider the dynamics between object detections [9]. In our paper, we exploit the dynamics of the robot between object detections to quantify the certainty of an object detection measurement.

In this paper, we propose a novel framework that simultaneously uses object classification to improve pose estimation and pose estimation to improve the certainty of measurements from object classifiers. In order to robustly include semantic measurements into the estimation formulation, we introduce a higher-layer estimation framework that parses the vehicle trajectory into object detection events, and uses Hidden Markov Model (HMM) algorithms to estimate the certainty of the object detection events. This computed certainty, which improves on the certainty guaranteed by typical classification algorithms, is also used to improve the robustness of our semantic estimation algorithm.

This paper is structured as follows: in Section II, we review the problem formulation and introduce the format of the semantic data. We show the maximum-likelihood formulation of the SLAM problem in Section III. An alternative formulation for including the semantic data into the factor graph is derived in Section IV. In Section V, we introduce switch variables to improve the robustness of our algorithm. We propose a higher-layer estimation framework modeled

*This work was in part supported by AeroVironment, Inc., Boeing, and Caltech's Center for Autonomous Systems and Technologies (CAST).

[1] Karena Cai is a graduate student in Control and Dynamical Systems at the California Institute of Technology kcai@caltech.edu

as an HMM in Section VI to further improve the robustness of the formulation. Finally, the algorithm is summarized in Section VII and simulation results are presented in Section VIII.

II. Problem Formulation

Consider the traditional localization and mapping algorithm, where the goal is to simultaneously estimate the vehicle poses $\mathcal{X} \triangleq \{x_t\}_{t=0}^{T}$ and the position of a set of landmarks in the environment denoted by $\mathcal{L} \triangleq \{l_m\}_{m=1}^{M}$, given a set of continuous measurements $\mathcal{Z}_{\mathrm{c}} \triangleq \{z_{\mathrm{c},t}\}_{t=0}^{T}$. The landmarks are features in the environment that can easily be recognized, and the continuous measurements are range measurements to those landmarks. Note that $x \in \mathbb{R}^n$, and M is the number of landmarks in the environment. For this model, we assume we have a set of odometry measurements given by $\mathcal{B} \triangleq \{b_t\}_{t=0}^{T}$ to approximate the vehicle dynamics, where b_t gives the vehicle translation and rotation between discrete-time points of the vehicle trajectory. These odometry measurements can be given by methods like the Iterative Closest Point (ICP) algorithm. This estimation problem is typically formulated as the following maximum-likelihood problem:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}} = \mathrm{argmax}_{\mathcal{X},\mathcal{L}} \log p(\mathcal{Z}_c|\mathcal{X},\mathcal{L}). \tag{1}$$

In our problem, we consider a vision-based object classification algorithm that can detect K different objects given by $\mathcal{O} \triangleq \{o_k\}_{k=1}^{K}$. We assume that we have an a priori map that defines the positions of each object o_k and the corresponding region R_k where the object can be detected. We define the set of semantic measurements as $\mathcal{Z}_s \triangleq \{z_{\mathrm{s},t}\}_{t=0}^{T}$, where $z_{\mathrm{s},t} \in \mathbb{R}^K$. The measurement corresponding to the object detector of object o_k can be represented as a binary variable $z^k_{\mathrm{s},t} \in \{1,0\}$, where a measurement of 1 indicates that the object o_k has been detected and 0 indicates that it has not. Each object detection measurement has a corresponding confusion matrix $C^k \in \mathbb{R}^{2\times 2}$ that captures the ratio of false positive and negative detections. Let the variable v^k be an indicator variable representing whether the object o_k is in the field of view of the camera and can be detected. The elements of the confusion matrix are defined as: $c^k_{iv^k} = p(z^k_{\mathrm{s},t} = i|v^k)$. We assume these error statistics can be computed offline. In the case of a perfect classification algorithm, the confusion matrix would be the identity.

The goal of this paper is to formulate an estimation algorithm that improves the robustness and accuracy of conventional localization and mapping algorithms by incorporating these semantic measurements.

III. Maximum-Likelihood Formulation with Semantic Data

The estimation problem with semantic measurements can be formulated as a maximum-likelihood problem. The formulation is given as follows:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}} = \mathrm{argmax}_{\mathcal{X},\mathcal{L}} \log p(\mathcal{Z}_c, \mathcal{Z}_{\mathrm{s}}|\mathcal{X},\mathcal{L}) \tag{2}$$

Assuming the semantic measurements are independent from the continuous range measurements, the maximization problem can be rewritten as the following minimization problem:

$$\begin{aligned}\hat{\mathcal{X}}, \hat{\mathcal{L}} = \mathrm{argmin}_{\mathcal{X},\mathcal{L}} \sum_{t=0}^{T}\sum_{k=1}^{K} -\log(p(z^k_{\mathrm{s},t}|x_t)) - \\ \log(p(\mathcal{Z}_c|\mathcal{X},\mathcal{L})) - \log(p(\mathcal{B}|\mathcal{X})),\end{aligned} \tag{3}$$

where the first term in the formulation corresponds to the likelihood function of the semantic measurements, and the second and third terms represent the nonlinear least-squares terms associated with the continuous measurements and odometry measurements respectively. The likelihood function for the semantic measurements from object detector k can be derived from the object detector's confusion matrix C^k as follows:

$$p(z^k_{\mathrm{s},t}|x_t) = \prod_{a=\{0,1\}} \prod_{b=\{0,1\}} c_{a,b}^{\mathbb{1}(z^k_{\mathrm{s},t}=a)\mathbb{1}(v^k_t=b)}. \tag{4}$$

This likelihood function selects each element of the confusion matrix based on the semantic measurement $z^k_{\mathrm{s},t}$ and v^k_t, which indicates whether the object is in the field of view. The indicator variable v^k_t is a function of the pose-estimate x_t and the rotation matrix associated with the camera.

Depending on whether the measurement $z^k_{\mathrm{s},t}$ is a 1 or a 0, the likelihood function derived in (16) takes on a different shape. The likelihood function shown in Fig. 1 assumes the object is in the field of view when inside the region R_k associated with object o_k. Under this assumption, we can see how the likelihood function promotes the region corresponding to the detected object when the measurement $z^k_{\mathrm{s},t} = 1$ and demotes the region when $z^k_{\mathrm{s},t} = 0$.

For the likelihood estimation formulation, where we solve for (3), we only include measurements where $z^k_{\mathrm{s},t} = 1$. The positive detection events are the measurements that give the most information, since they promote a very small region when they occur. Further, a measurement where $z^k_{\mathrm{s},t} = 0$ does not imply the vehicle is not in the region associated with the object. Instead, the vehicle could simply not have the object in its field of view, but still be in the region R_k.

Note that the likelihood function is a discrete, nonlinear function that must be approximated by a smooth function in order to be implemented in any factor-graph estimation algorithms like gtsam [11], which relies on gradient-based methods for solving the optimization problem. The details of this approximation are given in the Appendix.

Although this model improves the robustness of the estimation algorithm, the likelihood function does not take into consideration higher-level details about the measurements like their persistence over time. In the next section, we therefore introduce an alternative formulation for including object detection events as nonlinear factors that impose state constraints similar to loop closure detections.

IV. Factor-Graph Formulation with Semantic Data as State Constraints

In this section, we treat each object detection measurement $z^k_{\mathrm{s},t} = 1$ as a state constraint. We do not consider measurements when $z^k_{\mathrm{s},t} = 0$ in our factor-graph formulation for the same reasons we excluded them in the maximum-likelihood formulation. Since the standard factor-graph formulation does not allow for explicit state-constraints, we introduce a relaxation, and use the following nonlinear least-squares factor to represent the constraint imposed by a positive object detection measurement:

$$f(z^k_{\mathrm{s},t}, x_t) = z^k_{\mathrm{s},t} f^k_1(x_t). \tag{5}$$

This semantic factor is defined to reflect the same properties as the discrete likelihood function described in Section III. The comparison between the factors derived for the likelihood function and the factor derived here can be seen in Fig. 1. The factor $f_1(x_t)$ is defined as the following piecewise function:

$$f^k_1(x_t) = \begin{cases} 0 & d^k_h(x_t) = 0 \\ \alpha \exp(-\frac{\beta}{d^k_h(x_t)}) & d^k_h(x_t) > 0 \end{cases}, \tag{6}$$

where $d_h(x_t)$ is the shortest distance from x_t to the boundary of the region corresponding to object k given by R_k. Although the factor representing the likelihood function and the customized factor are similar, the customized factor improves the estimation accuracy further because of properties of its gradient.

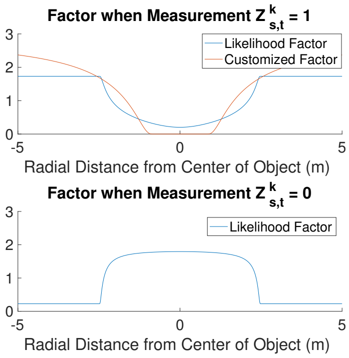

Fig. 1. The nonlinear least squares factors added to the graph corresponding the semantic measurement $z^k_{\mathrm{s},t} = 1$ is in the top figure and $z^k_{\mathrm{s},t} = 0$ on the bottom. The factor added corresponding to the negative log likelihood function described in (16) and the customized factors formed by piecewise inverse exponential functions correspond to the blue and red plots respectively. The parameter α for the inverse exponential factor is chosen to be 3.5 to approximate the same properties as the factor derived from the likelihood function.

The gradient of the function $f^k_1(x_t)$ is nonzero even when the estimate is far from the object detection region, so it still acts towards improving the estimate each time a positive object detection measurement occurs. Further, the parameters α and β can be modified to change the scale and rate of the inverse exponential functions respectively. With these additional features, the new formulation with semantic measurements becomes the following minimization problem:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}} = \mathrm{argmin}_{\mathcal{X},\mathcal{L}} \sum_{t=0}^{T}\sum_{k=1}^{K} \|f(z^k_{\mathrm{s},t}, x_t)\| - \log(p(\mathcal{Z}_c|\mathcal{X},\mathcal{L})) - \log(p(\mathcal{B}|\mathcal{X})). \tag{7}$$

False positive measurements will cause the wrong state-constraint factors to be imposed and will result in poor estimation results. In the next section, we introduce switchable constraints, taken from the loop-closure literature, to account for the possibility of bad measurements.

V. Robust Factor-Graph Formulation with Semantic Data and Switchable Constraints

In traditional SLAM algorithms, switchable constraints are introduced into the optimization formulation to improve the algorithm's performance when false positive data associations or loop-closure detections occur [7]. We propose the following addition of switchable constraints to improve the robustness of our algorithm to false semantic measurements:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}}, \hat{\Gamma} = \mathrm{argmin}_{\mathcal{X},\mathcal{L},\Gamma} \sum_{t=0}^{T}\sum_{k=1}^{K} \Big(\|\Psi(\gamma^k_t) f(z^k_{\mathrm{s},t}, x_t)\|_\Sigma + \|\gamma^k_t - \bar{\gamma}^k_t\|_\Lambda \Big) - \log(p(\mathcal{Z}_c|\mathcal{X},\mathcal{L})) - \log(p(\mathcal{B}|\mathcal{X})), \tag{8}$$

where $\Gamma \triangleq \{\gamma_t\}_{t=0}^T$ is the set of all switch variables, $\gamma_t \in \mathbb{R}^K$, and $\bar{\Gamma} \triangleq \{\bar{\gamma}_t\}_{t=0}^T$ is the set of priors on the switch variables. The variables Σ and Λ are optimization hyperparameters that determine the weight of the factors corresponding to the state constraints and the switch variable priors. The function $\Psi : \mathbb{R} \mapsto [0,1]$ is a function which takes a real value and maps it to the closed interval between 0 and 1. We choose $\Psi(\gamma^k_t) = \gamma^k_t$ to be a linear function of the switch variables and to constrain the switch variables between $0 \leq \gamma^k_t \leq 1$ since these choices have been empirically shown to work well [7].

Each switch variable γ^k_t quantifies the certainty of its associated semantic measurement $z^k_{\mathrm{s},t}$. When the switch variable γ^k_t is set to 0, the certainty in the measurement is extremely low, and the influence of the state-constraint factor associated with the measurement $z^k_{\mathrm{s},t}$ gets disregarded. Probabilistically, the switch variable is modifying the information matrix associated with the semantic factor such that $\hat{\Sigma}^{-1} = \Psi(\gamma^k_t)^2\Sigma^{-1}$ [7]. This means the covariance of the semantic measurement is unchanged from Σ when $\Psi(\gamma^k_t) = 1$ and the certainty in the measurement is high, but scales with $\Psi(\gamma^k_t)^2$ when $\Psi(\gamma^k_t) < 1$. For typical object classifiers, since the rate of false positive measurements is relatively low, we can default to trusting the measurements, so the switch priors $\bar{\gamma}^k_t$ are set to 1. Thus, both the certainty of each object detection measurement and the pose estimates are optimized for in this framework.

Although the formulation in (8) is much more robust to semantic measurement errors, setting the prior on all switch variables to 1 will sometimes cause the optimization to converge to the wrong solution. If we can compute the certainty of each object detection measurement by leveraging both the error statistics of the object classifier algorithms and the vehicle dynamics, we can construct a more accurate prior on the switch variables. In the next section, we propose a formulation where we use an HMM to compute the marginal probabilities of object detection events. These marginal probabilities can be used to set the prior on the switch variables in the optimization framework.

VI. HMM Formulation

Higher-level properties of the estimation formulation, like the persistence of semantic measurements over time and the relative vehicle dynamics, can be used to improve the certainty of object detection measurements. In this section, we propose a discrete-state estimation framework in the form of a Hidden Markov Model (HMM) that provides a measure on the certainty of the semantic measurements that occur during object detection events.

We consider a discrete-state representation of the vehicle trajectory in terms of object detection events. The vehicle trajectory can be parsed into different object detection events based on the persistence of semantic measurements in $\mathcal{Z}_\text{s}$ over time. The continuous-state estimation, which we refer to as the lower-layer estimation framework, occurs on the time-scale of t whereas the discrete-state estimation, which we refer to as the higher-layer estimation framework, occurs on the time-scale of τ. This is also shown more clearly in Fig. 3. Once an object detection event has been detected, we represent the detection event with a discrete state s_τ. This discrete state s_τ has a time interval in the continuous-state estimation time domain given by $t_\tau = [t_{\tau,i}, t_{\tau,f}]$ and a set of semantic measurements $Y_\tau \triangleq \{z_{\text{s},t}\}_{t=t_{\tau,i}}^{T_f = t_{\tau,f}}$ that occur over the time interval t_τ. The semantic measurements associated with the object o_k during this time interval are defined as: $Y_\tau^k \triangleq \{z_{\text{s},t}^k\}_{t=t_{\tau,i}}^{T_f = t_{\tau,f}}$.

The vehicle trajectory can then be represented as a sequence of states $\mathcal{S} \triangleq \{s_\tau\}_{\tau=1}^{Q}$, where Q is the number of object detection events that occur over the trajectory and $s_\tau = \{o_1, o_2, ..., o_k\}$. The notation $s_\tau = o_i$ means that object o_i has been seen during the detection event s_τ.

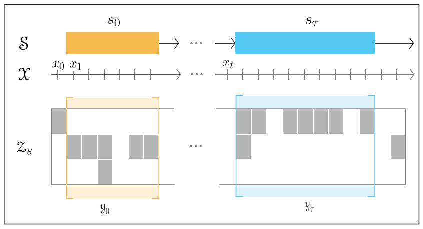

Fig. 2. This figure shows the relation between the discrete-state HMM and the continuous-state estimation. The semantic measurements $\mathcal{Z}_\text{s}$ are illustrated by the grid, where each row represents the measurements received by an object classifier. The grey boxes represent $z_{\text{s},t}^k = 1$ and the white boxes represent $z_{\text{s},t}^k = 0$. We can see that an object detection event, which is represented as a discrete state in the HMM, corresponds to a time interval in the continuous-state estimation framework.

The time-sequence of object detection events can be modeled as an HMM since the memoryless-Markov property holds, i.e. $p(s_\tau | s_{0:\tau-1}) = p(s_\tau | s_{\tau-1})$. The HMM estimation formulation and the algorithms for computing the switch prior, which are used to improve the optimization formulation in (8), are described in the following sections.

A. Parsing Trajectory into Object Detection Events

In this model, we assume only one object detection event occurs at a given time. The trajectory $\mathcal{X}$ can be parsed into different object detection events based on the semantic measurements $\mathcal{Z}_\text{s}$. Each object classifier is associated with a sequence of measurements given by $\{z_{\text{s},t}^k\}_{t=0}^{T}$ and a confusion matrix C^k that describes the algorithm's error statistics. In the event that the object o_k is visible, the frequency of nonzero measurements can be approximated by $C_{11}^k = p(z_{\text{s},t}^k = 1 | v_t^k = 1)$. We therefore define an object detection event for the object o_k as occurring when the proportion of nonzero measurements over a minimum time interval exceeds the threshold value set by $C_{11}^k - \varepsilon$. The value of ε is set to a value that depends on the certainty of the statistics given by the confusion matrix. The object detection event is terminated when the proportion of nonzero measurements decreases to less than the threshold value of $C_{11}^k - \varepsilon$.

B. Transition and Observation Matrices

HMMs are typically defined by a single transition matrix and a single observation matrix. The hybrid nature of our estimation formulation means continuous states have elapsed between the discrete states representing object detection events, and that a set of semantic measurements, denoted by Y_τ, have elapsed during each object detection event. Since we want to incorporate continuous-state pose estimates and semantic measurements into our HMM formulation, the transition and observation matrices are time-varying and dependent on the lower-layer estimates and measurements. In particular, the transition matrices are a function of the pose-estimates between discrete states representing object detection events, and the observation matrices are a function of Y_τ, the semantic measurements that have elapsed during the time interval t_τ corresponding to the discrete state s_τ. Each element of the transition matrix between discrete states $s_{\tau-1}$ and s_τ is defined as follows:

$$\begin{aligned} A(s_{\tau-1}, s_\tau)_{ij} &\triangleq p(s_\tau = o_j | s_{\tau-1} = o_i, \hat{x}_{t_{\tau,i}}, \hat{x}_{t_{\tau-1,f}}) \\ &\propto \exp(-\frac{1}{2} \| d_{ij} - \hat{d}_{\tau-1,\tau} \|), \end{aligned} \tag{9}$$

where $d_{ij} = \| p_i - p_j \|^{\frac{1}{2}}$, and p_i and p_j denote the positions of the center of mass (COM) of the objects o_i and o_j respectively. The distance $\hat{d}_{\tau-1,\tau} \triangleq \| \hat{x}_{t_{\tau,i}} - \hat{x}_{t_{\tau-1,f}} \|^{\frac{1}{2}}$ is the estimated distance traveled between object detection events. The rows of the transition probability matrix are normalized to sum to one. The transition probability is defined by the error between the actual distance of two objects from each other and the estimated distance traveled from one object detection event to another. The definition of the transition matrix would have to be modified to accommodate for the objects whose regions are not centralized around the objects' COM, because the distance traveled between object detection events could vary considerably. Examples of such objects include sidewalk or road detectors. This will be considered in future work.

Each element of the observation matrix for the discrete state s_τ is defined as follows:

$$\begin{aligned} O(\tau)_{ij} &\triangleq p(y_\tau = o_i, Y_\tau | s_\tau = o_j) \\ &= p(y_\tau = o_i | Y_\tau) p(Y_\tau | s_\tau = o_j). \end{aligned} \tag{10}$$

The probability $p(Y_\tau | s_\tau = o_j)$ is the likelihood of a sequence of semantic measurements over the time interval t_τ given that the object detection event corresponds to object o_j. When conditioned on $s_\tau = o_j$, each measurement in the sequence Y_τ^j can be modeled as a Bernoulli random variable with the probability of a nonzero measurement given by C_{11}^j. Thus, the probability $p(Y_\tau | s_\tau = o_j)$ can be approximated by how well the sequence of measurements Y_τ^j fits a Bernoulli distribution with parameter $p = C_{11}^j$. This probability can be computed with a Chi-Squared goodness of fit test [12].

Since there is no discrete-state observation of the system, we define y_τ to be a function of the sequence of measurements Y_τ. The discrete-state observation is the object that corresponds to the maximum likelihood for the sequence of semantic observations, which means $\bar{y}_\tau = \text{argmax}_{o_k} p(Y_\tau | s_\tau = o_k)$ for $k = 1, ..K$. Thus, the probability of a discrete-time measurement conditioned on the continuous-state observations becomes defined as follows:

$$p(y_\tau = o_i | Y_\tau) = \begin{cases} 1 & \text{if } o_i = \bar{y}_\tau \\ 0 & \text{if } o_i \neq \bar{y}_\tau \end{cases}. \tag{11}$$

Therefore, an observation matrix for each discrete-state s_τ can be derived for the HMM as a function of the sequence of semantic measurements Y_τ.

C. Computing Marginal Probabilities

The Viterbi algorithm can be used to determine the most probable sequence of states given a set of observations. We use a modified version of the Viterbi formulation given as follows:

$$\begin{aligned} \hat{\mathbb{S}} = \text{argmax}_{\mathbb{S}} \sum_{\tau=1}^{\tau,f} & \log(p(y_\tau, Y_\tau | s_\tau)) + \\ & \log(p(s_\tau | s_{\tau-1}, \hat{d}_{\tau-1,\tau})). \end{aligned} \tag{12}$$

This equation accounts for the dependencies of the transition and observation matrices on the time-varying lower-layer estimates and measurements. Once the most-probable sequence of states for the HMM are derived from the Viterbi algorithm in (12), the marginal probabilities $p(s_\tau = o_i | Y_{0:\tau,f})$ can be computed with dynamic programming using a modified version of the Forwards-Backwards algorithm. The variable $Y_{0:\tau,f}$ denotes all the measurement sequences corresponding to object detection events that have been observed. Details of this computation can be found in the Appendix.

The Forwards-Backwards algorithm therefore defines a certainty associated with the most-probable sequence of object detection events. The switch prior associated with the semantic measurements that occur over the time-intervals corresponding to these object detection events are computed in the following section.

D. Switch Prior Derivation

The switch prior γ_t^k associated with a semantic measurement $z_{\text{s},t}^k$ quantifies the reliableness of the measurement. When an object detection event corresponding to object o_k occurs, the marginal probability from the Forwards-Backwards algorithm gives us $p(s_\tau = o_k | Y_{0:\tau,f})$, which is a metric for the certainty that object o_k for the time interval t_τ associated with the detection event. This means that over the time interval t_τ, the measurements where $z_{\text{s},t}^k = 1$ should be proportional to the certainty of the detection event. Thus, we define the switch priors for the time interval t_τ where $s_\tau = o_k$ for every measurement for which $z_{\text{s},t}^k = 1$ as follows:

$$\gamma_t^k = p(s_\tau = o_k | Y_{0:\tau,f}, y_{0:\tau,f}) \tag{13}$$

In the case where the certainty in the object detection event $s_\tau = o_k$ is high, the switch priors corresponding to $z_{\text{s},t}^k = 1$ will be very close to 1.

The HMM formulation only gives a certainty metric for positive object detection events. The switch prior must also be derived for any semantic measurements where $z_{\text{s},t}^k = 1$ and the measurement does not occur during a time interval specified by an object detection event. These measurements occur during a non-detection event which occurs over the time interval in the continuous-state estimation time domain given by $t_{\tau\emptyset} = [t_{\tau\emptyset,i}, t_{\tau\emptyset,f}]$, and has semantic measurements given by $Y_{\tau\emptyset}^k = \{z_{\text{s},t}^k\}_{t=\tau\emptyset,i}^{T_f=\tau\emptyset,f}$. To compute the probability that all measurements during the time interval correspond to a non-detection event, which we define as $p(s_{\tau\emptyset} = \emptyset | Y_{\tau\emptyset}^k)$, we can compute how well the sequence of measurements in $Y_{\tau\emptyset}^k$ fits to a Bernoulli distribution with parameter C_{10}^k. The switch prior associated with the measurements outside the object detection are set to have a certainty proportional to $1 - p(s_{\tau\emptyset} = \emptyset | Y_{\tau\emptyset}^k)$. Thus, when the certainty that the non-object detection event has occurred is high, the switch priors corresponding to $z_{s,t}^k = 1$ will be very close to 0.

VII. System Architecture

In this section, we summarize the final estimation architecture. There are two estimation processes that are occurring on different time-scales: the continuous-state estimation process with switchable constraints and the discrete-state estimation of object detection events. Each process is iteratively improving the other, and the dependencies of the two processes can be seen in Fig. 3. The continuous-state estimation framework with switchable constraints is given by a modified version of the maximum-likelihood formulation.

The continuous-state optimization problem can be formulated as follows:

$$\hat{\mathcal{X}}, \hat{\mathcal{L}}, \hat{\Gamma} = \text{argmax}_{\mathcal{X},\mathcal{L},\Gamma} \log(p(\mathcal{Z}_\text{s}, \mathcal{Z}_c | \mathcal{X}, \mathcal{L}, \Gamma_0)) \tag{14}$$

The discrete-state estimation framework operates on the time scale of object detection events given by τ. Once an object-detection event has been classified, as described in Section VI-A, the pose-estimates and semantic measurements from (14) are used to derive the transition and observation matrices of the HMM. This HMM is used to model the discrete-state representation of the trajectory. The most probable sequence of object detection events, represented by the set of discrete states $\mathcal{S}$, is then solved as follows:

$$\hat{\mathcal{S}} = \text{argmax}_{\mathcal{S}} \log(p(\mathcal{Z}_s, \mathcal{Z}_c | \mathcal{S}, \hat{\mathcal{X}})) \tag{15}$$

The Forwards-Backwards algorithm is then used to compute the marginal probabilities associated with the maximum-likelihood sequence of object detection events.

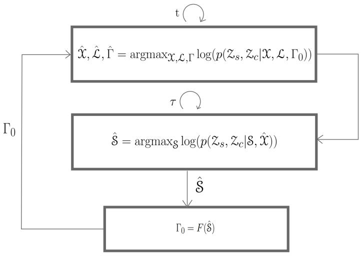
t
$\hat{\mathcal{X}}, \hat{\mathcal{L}}, \hat{\Gamma} = \text{argmax}_{\mathcal{X},\mathcal{L},\Gamma} \log(p(\mathcal{Z}_s, \mathcal{Z}_c | \mathcal{X}, \mathcal{L}, \Gamma_0))$
$\hat{\mathcal{X}}$
τ
Γ_0
$\hat{\mathcal{S}} = \text{argmax}_{\mathcal{S}} \log(p(\mathcal{Z}_s, \mathcal{Z}_c | \mathcal{S}, \hat{\mathcal{X}}))$
$\hat{\mathcal{S}}$
$\Gamma_0 = F(\hat{\mathcal{S}})$

Fig. 3. The system architecture is comprised of two layers: the lower layer shown in the top box represents the factor-graph formulation with switchable constraints and updates at every time step t, whereas the higher layer shown in the bottom box represents the HMM estimation framework, and computes switch prior variables after every object detection event τ. The switch variables are denoted by Γ and the switch prior is given by Γ_0.

These marginal probabilities are used to compute priors on the switch variables associated with the semantic measurements in (14). Therefore, the higher and lower-layer estimation processes are simultaneously improving the pose estimate and the certainty of object detection events.

VIII. Simulation Results

We investigate the performance of our algorithms in simulation. We consider an object classifier that can detect three different objects, each of which is associated with a known radial region in a 2-D map that is shown in Fig. 4, and each of which has a confusion matrix whose parameters are defined in Table I.

TABLE I
Confusion Matrix Parameters

	o_1	o_2	o_3
$p(z^k_{s,t} = 1 \| v^k)$	0.02	0.03	0.05
$p(z^k_{s,t} = 0 \| v^k)$	0.2	0.15	0.1

There are four landmarks that provide the vehicle with range measurements when detected. The data association problem for the landmarks are assumed to be solved in this formulation. The vehicle traverses an ellipsoid trajectory and goes through each of the object detection regions during its path. We introduce noisy odometry measurements in our simulation. The range measurements to the landmarks mitigate the estimation error when included in the traditional SLAM algorithms. We investigate how the introduction of semantic measurements improves the estimation even further.

In our simulations, we run three different algorithms to estimate the vehicle trajectory. First, we run the algorithm with range measurements to the four different landmarks, which is what is conventionally available in the SLAM community. Second, we use the maximum-likelihood formulation with the smoothed likelihood function described in Section III. Finally, we test the two-layer estimation framework with the switch variables and the priors from the HMM. In our simulations, the estimation frameworks are implemented using gtsam, a factor-graph formulation commonly used for solving pose-graph estimation problems [13], [11].

For the different formulations, the noise model must be chosen for the factors corresponding to the state constraints imposed by the semantic measurements. In the algorithm involving the likelihood function, we use the identity matrix to define the covariance on the likelihood factor so that the Bayesian representation of the likelihood function is preserved.

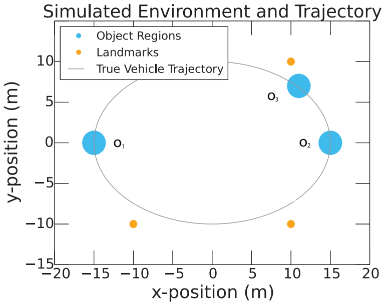

Fig. 4. The objects and their corresponding regions of detections R_k are shown in blue and the positions of the landmarks that are visible during the vehicle trajectory are shown in orange. The gray line represents the true vehicle path.

For the formulation with switch variables, the noise on the prior for the switch variables, which is given by Λ in (8) is chosen to be 0.01 when the switch prior values are chosen by the HMM and 10 when the switch prior is set to the default value of 1. This choice reflects our increase in certainty on the switch prior when we use the HMM formulation. The noise on the inverse-exponential factors, which is represented by Σ in (8) is chosen to be 0.5.

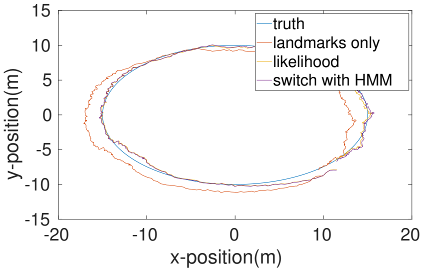

Fig. 5. The true trajectory can be compared to the estimated trajectory using three different algorithms. The algorithms that incorporate the semantic measurements improve the estimate significantly.

The estimated trajectory resulting from the different estimation algorithms are shown in Fig. 5. The squared error between the estimated trajectories and the true trajectory is shown in Fig. 6. We can see that the two algorithms that incorporate semantic measurements outperform the traditional

SLAM algorithm. We also see that the estimation framework with the smooth approximation of the likelihood function and the estimation framework with the switch variables and HMM-derived switch priors converge to very similar local minima.

Since different noise on the odometry measurements will contribute to different estimation results, we compare our estimation algorithms on a set of randomly generated odometry measurements. This way, we can evaluate the performance of the different algorithms over many different trials.

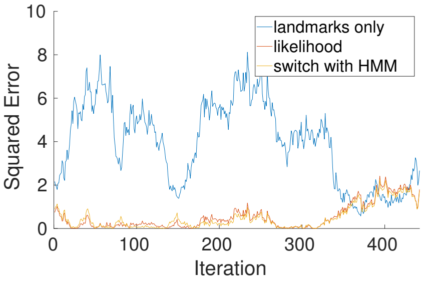

Fig. 6. The squared error between the true trajectory and the estimated trajectories from the three different estimation estimation algorithms.

The comparison of the different estimation algorithms is captured in Fig. 7 and Fig. 8. We see how the mean of the squared error over the entire trajectory for all the trials is notably smaller when the semantic measurements using either the likelihood algorithm or the two-layer algorithm.

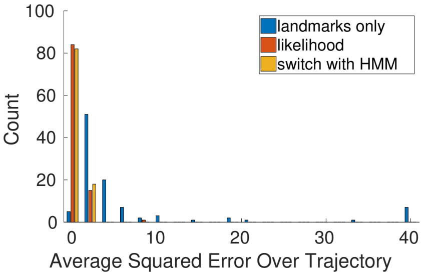

Fig. 7. The mean squared error over the trajectory path is computed for each of the different algorithms and the results from the trials can be compared in this histogram. The likelihood estimation algorithm and the two-layer estimation algorithm perform similarly, and both perform significantly better than the algorithm without semantic measurements.

The estimation framework with the likelihood function has more trials with very low average-squared errors but is less consistent than the two-layer estimation framework, since its distribution has higher variance. If we look at the squared error of the final position estimate, which is the value we are iteratively estimating, Fig. 8 shows that the HMM algorithm performs marginally better than the likelihood algorithm.

One of the advantages of using the HMM formulation is to guarantee a higher certainty for the measurements corresponding to object detection events by leveraging the persistence of the semantic measurements as well as vehicle dynamics. In this simulation, the sequence of events corresponding to the maximum likelihood given by the Viterbi algorithm was given as follows: $s_0 = o_2$, $s_1 = o_3$, and $s_2 = o_1$, meaning object 2 as detected first, followed by object 3 and then object 1. The corresponding marginal probabilities of these object detection events are $p(s_0 = o_2|Y_0) = 0.943$, $p(s_1 = o_3|Y_1) = 0.994$, and $p(s_2 = o_1|Y_2) = 0.997$. This shows certainty levels of object detection events that are much greater than the accuracy guaranteed by the c^k_{11} element in the confusion matrix for each object detector which were approximately 0.8 for each of the object classifiers in our simulation.

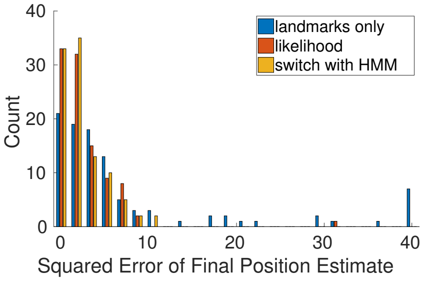

Fig. 8. For every trial corresponding to different odometry noise measurements, we use the three algorithms to estimate the trajectory. The mean squared error of the final position estimate is computed. The final position estimate is marginally better using the two-layer estimation framework.

The effectiveness of semantic measurements will depend on the frequency at which these objects are detected, but when objects have been detected, this estimation architectures proposed in this paper provide a robust way to incorporate these semantic measurements.

IX. Conclusion and Future Work

In this work, we introduce a robust estimation framework for incorporating probabilistic binary measurements, which are used to model data from vision-based object classification algorithms. We first introduce a formulation for solving a maximum-likelihood problem with the semantic measurement likelihood function modeled after the confusion matrix of object classifiers. We then derive a two-part estimation framework where the lower-layer is formulated as a factor-graph estimation problem, with each measurement corresponding to a state-constraining factor modeled after the discrete likelihood function and a switchable constraint. We also present a higher-layer estimation framework that takes into account measurement persistence and vehicle dynamics to compute the certainty of sets of semantic measurements corresponding to object detection events. These certainties capture which measurements are false positives, and are used to compute the switch priors in the lower-layer estimation algorithm. The advantage of including the higher-layer estimation framework is demonstrated in the presented numerical simulation. We show in simulation how the addition of semantic measurements in this framework improves the robustness and accuracy of the estimated trajectory.

In future work, we will model the probabilistic likelihood factors such that they account for a lower detection probability with increasing distance from the detected object. We will investigate extensions to include 3-D object data that incorporates height information, and test such algorithms on aerial vehicles. Finally, we also hope to incorporate object detection measurements when an a-priori map of objects is not known, but relative relations among different objects are known.

APPENDIX

A. Discrete Likelihood Function

The discrete likelihood function in Section II is a nonlinear discrete function. There are two separate likelihood functions depending on whether the semantic measurement $z^k_{\mathrm{s},t}$ is 1 or 0. We want to give a smooth approximation for the negative log of the discrete likelihood function. For this paper, we approximate the likelihood function with a bump function parameterized by the probabilities a and b with the following form:

$$f(r,a,b) = \sqrt{(k\exp(-\frac{1}{r_0^2 - r^2}) - \log(b))}, \tag{16}$$

where $k = \frac{\log(b)-\log(a)}{\exp(-\frac{1}{r_0^2})}$ and $r = \sqrt{(x_1^2 + x_2^2)}$, where x_1 and x_2 are the first and second coordinate of the state x_t. For the bump function corresponding to $z^k_{\mathrm{s},t} = 1$, the bump function parameters are set so $a = C^k_{11}$ and $b = C^k_{10}$ and for the measurement $z^k_{\mathrm{s},t} = 0$, the parameters are set so $a = C^k_{01}$ and $b = C^k_{00}$.

The nonlinear factor in a factor graph becomes the term $f(r,a,b)R^{-1}f(r,a,b)$, where R is the covariance associated with the measurement. This is why the square root is necessary in order to preserve the Bayesian representation of the likelihood function, and the covariance matrix associated with this factor is chosen to be the identity matrix.

B. Forwards-Backwards Algorithm

The modified Forwards-Backwards algorithm that can be used to accommodate the time-varying transition and observation matrices can be written as follows:

$$P(s_\tau = o_i, Y_{0:\tau,f}) = \frac{\alpha_{o_i}(\tau)\beta_{o_i}(\tau)}{\sum_{o_i'} \alpha_{o_i'}(\tau)\beta_{o_i}(\tau)}, \tag{17}$$

where the values $\alpha_{o_i}(\tau)$ and $\beta_{o_i}(\tau)$ are recursively defined below:

$$\alpha_{o_i}(\tau) = p_\tau(y_\tau|o_i)\sum_{o_i'} \alpha_{o_i'}(\tau-1)p_{\tau-1,\tau}(o_i|o_i') \tag{18}$$

and

$$\beta_{o_i}(\tau) = \sum_{o_i'} \beta_{o_i'}(\tau+1)p_{\tau-1,\tau}(o_i'|o_i)p_{\tau+1}(y_{\tau+1}|o_i'). \tag{19}$$

The dependencies of the transition and observation matrices on the object detection events are denoted by the subscripts of the probabilities in the above equation.

ACKNOWLEDGMENT

The authors would like to acknowledge Andrew Stuart for helpful conversations on data assimilation methods for this paper.

REFERENCES

[1] R. Ivanov, N. Atanasov, M. Pajic, G. Pappas, and I. Lee, "Robust estimation using context-aware filtering," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 590–597, Sept 2015.

[2] R. Ivanov, N. Atanasov, M. Pajic, J. Weimer, G. J. Pappas, and I. Lee, "Continuous estimation using context-dependent discrete measurements," *IEEE Transactions on Automatic Control*, pp. 1–1, 2018.

[3] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas, "Semantic localization via the matrix permanent," in *Proc. of Robotics and Science Systems*, 2014.

[4] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1722–1729, May 2017.

[5] A. Glover, W. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth, "Openfabmap: An open source toolbox for appearance-based loop closure detection," in *2012 IEEE International Conference on Robotics and Automation*, pp. 4730–4735, May 2012.

[6] L. Carlone, A. Censi, and F. Dellaert, "Selecting good measurements via l1relaxation: A convex approach for robust estimation over graphs," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2667–2674, Sept 2014.

[7] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph slam," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1879–1884, Oct 2012.

[8] P. Agarwal, G. D. Tipaldi, L. Spinello, C. Stachniss, and W. Burgard, "Robust map optimization using dynamic covariance scaling," in *2013 IEEE International Conference on Robotics and Automation*, pp. 62–69, May 2013.

[9] P. Robbel and D. Roy, "Exploiting feature dynamics for active object recognition," in *2010 11th International Conference on Control Automation Robotics Vision*, pp. 2102–2108, Dec 2010.

[10] T. Patten, M. Zillich, R. Fitch, M. Vincze, and S. Sukkarieh, "Viewpoint evaluation for online 3-d active object classification," *IEEE Robotics and Automation Letters*, vol. 1, pp. 73–81, Jan 2016.

[11] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "isam2: Incremental smoothing and mapping using the bayes tree," *International Journal of Robotics Resesarch*, vol. 31(2), pp. 216–235, 2011.

[12] N. Balakrishnan, V. Voinov, and M. S. Nikulin, *Chi-Squared Goodness of Fit Tests with Applications*. Academic Press, 2013.

[13] F. Dellaert and M. Kaess, "Factor graphs for robot perception," vol. 6, pp. 1–139, 01 2017.

[14] J. Civera, D. Gálvex-López, L. Riazuelo, J. D. Tardós, and J. M. M. Montiel, "Towards semantic slam using a monocular camera," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1277–1284, Sept 2011.

[15] M. W. Hofbaur and B. C. Williams, "Mode estimation of probabilistic hybrid systems," in *International Workshop on Hybrid Systems: Computation and Control*, 2002.

[16] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, pp. 1309–1332, Dec 2016.

[17] S. M. Oh, S. Tariq, B. N. Walker, and F. Dellaert, "Map-based priors for localization," in *IEEE Conf. on Intelligent Robots and Systems*, 2004.

[18] S. Choudhary, L. Carlone, C. Nieto, J. Rogers, Z. Liu, H. I. Christensen, and F. Dellaert, "Multi robot object-based slam," in *2016 International Symposium on Experimental Robotics*, pp. 729–741, Springer International Publishing, 2017.

[19] S. Ekvall, P. Jensfelt, and D. Kragic, "Integrating active mobile robot object recognition and slam in natural environments," in *IEEE Intl. Conf. on Intelligent Robots and Systems*, 2006.

[20] N. Atanasov, B. Sankaran, J. L. Ny, G. J. Pappas, and K. Daniilidis, "Nonmyopic view planning for active object classification and pose estimation," *IEEE Transactions on Robotics*, vol. 30, pp. 1078–1090, Oct 2014.

[21] Torralba, Murphy, Freeman, and Rubin, "Context-based vision system for place and object recognition," in *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 273–280 vol.1, Oct 2003.